c246 Problem Set 6

Due 15 March 2002 by email to homework@c246.lbl.gov

1) (15 points) The MEME program at http://meme.sdsc.edu offers three different options for how you think the motif might be distributed in the sequence. What are they, what do they mean, and, how do you think these alter the implementation of the version of the MEME algorithm discussed in class.

2) (15 points) Write a simple program that generates random DNA sequences of chosen length **N** defined by a two-component mixture model (as per MEME). The parameters the width of the motif **w**, a position weight matrix $\boldsymbol{\theta}_0$ describing the motif, a vector $\boldsymbol{\theta}_1$ describing the probabilities of finding a given base in background sequence and a parameter $\boldsymbol{\lambda}$ describing the probability that a motif begins at a given base [when you progressing along your sequence, at each position us $\boldsymbol{\lambda}$ to decide whether to spit out a motif of width w, or spit out a background base and go on].

3) (15 points) Using the program from problem 2 and reasonable values of **w**, $\boldsymbol{\theta}_0$, and $\boldsymbol{\theta}_1$ (pick a motif from the literature or get on from a database such as TRANSFAC at http://transfac.gbf.de/TRANSFAC/), generate sequences of length 2000 using the following series of values for $\boldsymbol{\lambda}$: 0.001, 0.005, 0.01, 0.05 and 0.1. Using the MEME website (or a local implementation of MEME if you have one), run MEME (using the two-component mixture model) on these sequences. How similar are the results for the different input sequences? What is different? How do the results change with the values of $\boldsymbol{\lambda}$.

4) (15 points) For DNA searches, the MEME website allows to you limit your search to only find palindromes in DNA sequences. Why is this restricted to DNA searches? How would you implement this constraint?

5) (10 points) The background model used by MEME and the Gibbs sampler described in your reading is rather primitive. Describe how it could be improved and how you would implement these improvements.

6) (15 points) Design an HMM to recognize trans-membrane domains in protein sequences. Describe both the architecture of the HMM and how you would go about estimating the parameters of the model.

7) (15 points) Design an HMM to find motifs of the sort detected by MEME.